

Les conventions sans la connaissance commune ?

Paul Egré* et Olivier Roy[#]

(Institut Jean-Nicod/ Munich Center for Mathematical Philosophy)

I. Introduction

Premier livre publié de David Lewis, *Convention* est aujourd'hui l'un des traités philosophiques les plus cités sur les fondements de la rationalité collective. La postérité de l'ouvrage s'étend bien au-delà du cercle des philosophes, elle rayonne jusque parmi les économistes, les psychologues et les linguistes. Dans ce livre, Lewis s'inspire notamment des travaux du théoricien des jeux Thomas Schelling pour proposer une définition originale de la notion de convention, fondée de façon pragmatique sur le concept d'équilibre stratégique au sein d'un jeu de coordination¹. Ce faisant, Lewis ne se contente pas d'emprunter à la théorie des jeux de son époque : il invente en particulier deux concepts qui connaîtront plusieurs années plus tard une fortune considérable en théorie des jeux, à savoir le concept de *connaissance commune*, et le concept de *jeu de signal*.

* Paul Egré est chargé de recherche au CNRS et membre de l'Institut Jean-Nicod (CNRS-EHESS-ENS). Ses principaux domaines de recherche sont la logique philosophique, la philosophie de la connaissance, et la philosophie du langage. En philosophie de la connaissance, il s'est intéressé en particulier au thème de la métacognition et des attitudes réflexives (dernier article paru : « Metacognitive perspectives on unawareness and uncertainty », en collaboration avec D. Bonnay, in M. Beran, J. Brandl, J. Perner and J. Proust (éds.), *Foundations of Metacognition*, 2012). Dans celui de la logique et de la philosophie du langage, ses travaux les plus récents portent sur les paradoxes du vague, thème sur lequel il a dirigé un ouvrage (*Vagueness and Language Use*, avec N. Klinedinst, Palgrave Macmillan, 2011) et publié plusieurs articles (dernier article paru : « Tolerant, Classical, Strict », en collaboration avec P. Cobreros, D. Ripley et R. van Rooij, *Journal of Philosophical Logic*, 41 (2), 347-395, 2012).

[#] Olivier Roy est Assistant Professeur au Munich Center for Mathematical Philosophy. Il a obtenu son doctorat à l'Institute for Logic, Language and Computation (Université d'Amsterdam) en 2008, sous la supervision de Johan van Benthem et de Martin van Hees. Ces travaux récents portent sur la théorie épistémique des jeux, la logique épistémique, la philosophie de l'action et la théorie de la rationalité. Parmi ses publications récentes on compte, « Substantive Assumptions in Interaction : A Logical Perspective » (avec Eric Pacuit, à paraître dans *Synthese*), « Agreement Theorems in Dynamic Epistemic Logic » (avec Cédric Dégremon, *Journal of Philosophical Logic*, Volume 41, Numéro 4 2012), « Towards a Theory of Play : A Logical Perspective on Games and Interaction » (avec Johan van Benthem et Eric Pacuit, *Games*, 2(1), 52-86, 2011) et « Interpersonal Coordination and Epistemic Support for Intentions with We-Content » (*Economics and Philosophy*, 26(3), 345-367, 2010).

¹ Voir Schelling (1960).

Le premier, le concept de connaissance commune, sera réintroduit et formalisé par l'économiste Robert Aumann en 1976, dans son article «Agreeing to Disagree»². Le second, le concept de jeu de signal, sera introduit indépendamment par l'économiste Michael Spence au début des années 1970 et, dans une version plus proche de ce qui intéressait Lewis, par les économistes Vincent Crawford et Joel Sobel au début des années 1980. Chacun de ces concepts sert un but différent dans l'ouvrage de Lewis. La notion de connaissance commune, introduite au chapitre 3, intervient dans la définition de la notion générique de convention que retient Lewis. La notion de jeu de signal, qui occupe les chapitres 4 et suivants, vise à rendre compte de façon naturaliste de la notion plus spécifique de convention linguistique, et par là de la signification des mots au sein d'un langage³.

Par elle-même, à savoir indépendamment des applications auxquelles elle donne lieu, la définition d'un jeu de signal ne prête guère à discussion. Pour Lewis, un jeu de signal est un jeu d'un type particulier, dans lequel l'un des joueurs, l'émetteur, possède une information sur le monde dont ne dispose par l'autre, le receveur. En fonction de ce qu'il observe, l'émetteur envoie un signal au receveur, lequel choisit d'agir de telle ou telle manière. Cette définition pose avant tout d'un cadre théorique, certes remarquablement fécond, mais dont toute la richesse dépend des hypothèses faites sur la structure du jeu (voir Skyrms 1996, 2010). La notion de connaissance commune, en revanche, est une notion qui continue de faire énigme, et dont différentes définitions ont été proposées depuis la définition pionnière de Lewis (pour un panorama critique, voir en particulier Vanderschraff et Sillari 2009 et Paternotte 2011). C'est cette notion que nous proposons d'examiner ici, remettant une discussion de la contribution de Lewis sur les jeux de signal à une autre occasion⁴.

Concernant la connaissance commune, deux problèmes nous semblent mériter une attention particulière. Le premier concerne la définition de la notion de connaissance commune et la question de sa réalité psychologique.

² Sur la généalogie du lien entre la définition de Lewis et celle d'Aumann, voir l'entretien avec R. Aumann réalisé par O. Roy in Hendricks et Roy (2010). Sur la généalogie du concept de connaissance commune chez Lewis lui-même, l'une des influences manifestes est H.P. Grice (1957), amplement discuté dans *Convention*. Le concept apparaît par ailleurs au même moment chez Schiffer (1972), qui parle de « connaissance mutuelle ». Schiffer (communication personnelle) nous a confié que Lewis et lui avaient correspondu à cette époque, l'un et l'autre reconnaissant une double paternité de la notion.

³ Sur le lien entre connaissance commune et convention linguistique, le texte le plus explicite de Lewis après *Convention* est par ailleurs Lewis (1973).

⁴ Sur les liens entre convention et jeux de signal, nous renvoyons au récent article de D. Rothschild (2012).

On présente couramment la notion de connaissance commune comme suit : chacun des agents au sein d'un groupe sait que *P*, chacun sait que chacun sait que *P*, et ainsi de suite, *ad infinitum*. La définition originale de Lewis est plus indirecte, comme nous le verrons, et plus complexe, et il vaut la peine de comprendre pourquoi. Le second problème concerne la pertinence de la notion, en l'occurrence le lien envisagé par Lewis entre la notion de connaissance commune et la notion de coordination entre les agents. Pour Lewis, une convention est essentiellement une régularité dans le comportement des agents, qu'on peut décrire comme un équilibre stratégique dans un jeu de coordination. Pour Lewis, cependant, cet équilibre n'aura le statut d'une convention que s'il est fondé sur un système d'attentes et de croyances mutuelles qui est connaissance commune entre les agents. Mais qu'en est-il véritablement ? Pour qu'un équilibre de coordination ait lieu entre des agents, et que celui-ci accède au statut d'une régularité de comportement, est-il toujours requis d'avoir des croyances sur les croyances d'autrui, et est-il nécessaire d'avoir une hiérarchie de croyances aussi forte que celle impliquée par la notion de connaissance commune⁵ ?

L'insistance mise par Lewis sur la notion de connaissance commune semble indiquer une réponse positive à la question. Toutefois, un examen plus détaillé invite au scepticisme. Ce scepticisme, manifesté notamment par Skyrms, jette le doute sur la pertinence même de la notion de connaissance commune. Pour Skyrms, en effet, les principes de rationalité qui gouvernent la coordination entre les agents pourraient fort bien n'être que des principes de plus bas niveau, ne requérant pas d'avoir des croyances sur les croyances d'autrui (Skyrms 2004). Le but de notre contribution est par conséquent d'éclaircir la question du rôle de la connaissance commune dans la coordination, dans le souci de mieux cerner le rôle que jouent les attentes d'ordre supérieur dans la rationalité collective et quant au fondement des conventions. Pour cela, notre démarche consiste à mettre en regard le texte de Lewis avec les résultats plus récents de la théorie épistémique des jeux (elle-même issue des travaux de Aumann), visant à faire apparaître à quelles conditions sur les croyances des agents un équilibre est atteignable dans un jeu. Ces résultats, comme nous le verrons, laissent la place à des interprétations divergentes sur le rôle que peut jouer la

⁵ On notera que la notion de *connaissance* commune est définie en terme d'une hiérarchie de *croyances*. Ceci est fidèle à la définition qu'en donne Lewis (voir ci-après Section 5). Nous signalerons toutefois, notamment dans la Section 6, les cas dans lesquels la distinction entre croyance et connaissance proprement dit importe pour les résultats présentés.

connaissance commune pour la réalisation des équilibres. Néanmoins, nous verrons en quel sens on peut les rapprocher de la conception même que Lewis se faisait du rôle de la connaissance commune dans les conventions.

II. Équilibre stratégique et convention

Un problème de coordination est défini par Lewis comme une situation dans laquelle au moins deux agents ont à choisir une action de façon indépendante, et de telle manière que les intérêts des agents sont identiques quelle que soit l'issue du jeu. Comme exemple typique, considérons l'exemple de la conduite automobile, discuté par Lewis. Imaginons un pays dans lequel il n'existe encore aucune règle explicite de conduite automobile (une situation rousseauiste ou encore humienne, telle que la caractérise Lewis). Deux agents approchent chacun d'un virage en sens opposé. Chacun peut choisir soit de rouler à droite, soit de rouler à gauche. Si les deux agents roulent chacun à droite, ou s'ils roulent chacun à gauche, ils poursuivent leur route sans encombre. Si l'un choisit de rouler à droite et l'autre à gauche, les agents n'auront pas le temps de dévier, et ils auront un accident. Le problème peut être représenté par la matrice de jeu suivante :

	G	D
G	1, 1	0,0
D	0,0	1, 1

Dans cet exemple, les gains marqués à gauche représentent ceux du joueur 1 (qui choisit une action Droite ou Gauche en rangée), et ceux à droite celui du joueur 2 (qui choisit une action Droite ou Gauche en colonne). Un gain de 1 (passage sans encombre) est préférable pour chaque agent à un gain de 0 (accident). Comme l'illustre la matrice, les gains sont identiques pour chaque agent en chaque issue du jeu. Ce jeu comporte en outre deux équilibres stratégiques (ou *équilibres de Nash*), à savoir les actions (D, D) et (G, G). Le profil d'action (D, D) constitue un équilibre car étant donnée l'action D du joueur 1, le joueur 2 n'a pas de meilleure action que de jouer D, et réciproquement, étant donnée l'action du joueur 2 de jouer D, le joueur 1 n'a pas de meilleure action que de jouer D. Un équilibre de Nash, plus généralement, est un profil d'actions (de stratégies) tel qu'aucun des joueurs n'a intérêt à en dévier de façon unilatérale. Comme l'indique cet exemple, la coordination entre les agents est donc réussie si les

agents choisissent des actions à l'équilibre. Notons qu'un problème de coordination n'implique pas que les agents aient nécessairement les mêmes actions à leur disposition, ni, quand c'est le cas, que l'équilibre soit toujours une paire d'actions identiques, comme dans l'exemple de la conduite.

L'observation que fait Lewis en préambule de sa démarche est que chaque équilibre stratégique dans le jeu de la conduite constitue le *point de départ* possible d'une convention. Vue de façon normative, une convention est une règle du type : « il faut rouler à gauche », ou « il faut rouler à droite ». Mais la démarche de Lewis est d'envisager de telles règles normatives à partir des comportements qui les ont précédées. Selon Lewis, pour aboutir à une convention, il faut d'abord qu'il y ait eu coordination effective entre les agents. C'est rarement une prescription explicite qui peut donner lieu à une convention, c'est plutôt la coordination effective et tacite qui en constitue le fondement. C'est pourquoi, plutôt que de parler de *règle*, Lewis caractérise d'abord les conventions de façon descriptive comme des *régularités* de comportement.

Afin d'en venir au problème de la connaissance commune, trois remarques méritent d'être faites sur les conventions ainsi comprises. Premièrement, une convention, relativement à un problème de coordination, implique pour Lewis une part d'arbitraire, et donc de choix. Ici, l'arbitraire et le choix proviennent de l'existence d'équilibres multiples dans le jeu. Rien dans notre exemple n'implique pour les agents qu'il soit plus rationnel de rouler à droite que de rouler à gauche. Cet arbitraire ne signifie pas pour autant que la convention soit immotivée. En l'occurrence, le principe commun aux deux équilibres (D, D) et (G, G) pourrait s'exprimer (de façon intensionnelle plutôt qu'extensionnelle) comme : « rouler du même côté de la route (que les autres) ».

Deuxièmement, il ne suffit pas d'atteindre un équilibre dans un jeu de coordination pour que cet équilibre constitue une convention. Comme nous l'avons déjà mentionné, l'atteinte de l'équilibre est pour Lewis le point de départ possible, mais n'est pas à lui seul suffisant pour qu'une régularité comportementale mérite le titre de convention. Imaginons une population d'agents qui jouent au même jeu que précédemment, mais qui doivent renouveler leur choix à chaque virage. Le fait que deux agents aient roulé à droite chacun à un premier virage ne fait pas de cet équilibre réalisé en un coup une convention. Il se pourrait qu'au prochain coup, la bonne action à réaliser pour chacun soit de rouler à gauche, si jamais les automobilistes qu'ils croisent choisissent cette fois de rouler à gauche. Être un équilibre est donc une condition nécessaire à un profil d'action pour constituer une

convention, mais ce n'est pas une condition suffisante.

Ce point nous amène naturellement à notre troisième remarque. Si l'équilibre n'est pas une condition suffisante pour qu'un comportement soit conventionnel, quelles conditions faut-il y ajouter ? C'est à ce niveau précisément qu'intervient chez Lewis la notion d'anticipation ou d'attente d'un agent relativement au comportement d'autrui. En effet, la première définition de la notion de convention à laquelle aboutit Lewis (1969, p. 42) dit qu'une régularité de comportement R au sein d'une population P d'agents dans une situation récurrente S est une convention si et seulement si trois conditions sont réunies :

- (1) chacun se conforme à R
- (2) chacun s'attend à ce que chaque autre se conforme à R
- (3) chacun préfère se conformer à R à la condition que les autres le fassent

Dans cette définition, si R est vu comme un profil d'actions dans un jeu de coordination (ou comme l'action pertinente à chaque joueur relativement à ce profil), on voit que les conditions (1) et (3) correspondent simplement à la caractérisation de l'équilibre de Nash. La différence vient de la condition (2). Dans notre exemple, supposé que (D, D) soit l'équilibre conventionnel, cette condition dit que le joueur 1 a des raisons de croire que le joueur 2 se conformera à R = (D, D), et idem que le joueur 2 a des raisons de croire que le joueur 1 se conformera à R = (D, D). Incidemment, nous avons reformulé ici la notion d'anticipation (ou d'attente, *expectation* en anglais) en terme de *raison de croire*, car Lewis substitue cette dernière à la notion d'anticipation dans la définition qu'il donne de la connaissance commune au chapitre suivant du livre. La condition (2) de la définition (préliminaire) de la notion de convention par Lewis implique donc que, pour qu'il y ait convention, il faut que les agents aient des croyances sur le comportement d'autrui. Cette condition semble plausible, car sans de telles anticipations sur le comportement d'autrui, il semble qu'il ne puisse y avoir que des équilibres fortuits et transitoires au cours des situations récurrentes S, mais en aucun cas une régularité de comportement.

Lewis, cependant, ne se contente pas de cette première définition de la notion de convention. Au chapitre suivant (1969, p. 58), Lewis propose un amendement considérable. Il maintient les conditions (1) à (3) précédentes, mais demande en outre que, pour que R soit une convention, il faut et il suffit qu'il soit *connaissance commune* parmi les agents de la population P

que les conditions (1) à (3) sont réunies, ce qui implique que chacun sache ces conditions, sache que chacun les sait, et ainsi de suite. À elles seules, les conditions (1)-(3) de la première définition d'une convention ne requièrent que des croyances de premier ordre chez les agents relativement au profil d'actions R. En passant à la seconde définition de la notion de convention, Lewis impose une condition beaucoup plus forte, qui demande que les agents aient des croyances sur les croyances les uns des autres, et des croyances de niveau encore supérieur. En quoi ce renforcement de la définition serait-il nécessaire ?

III. Les croyances d'ordre supérieur sont-elles nécessaires ?

Pour y répondre, considérons à nouveau l'exemple de la conduite et demandons-nous ce qui pourrait justifier que les conventions dépendent de croyances d'ordre supérieur.

Supposons que le joueur 1 et le joueur 2 se sont croisés une fois à un virage et que chacun roulait à droite, et supposons en outre qu'ils doivent se croiser à nouveau à ce même virage. Imaginons que le joueur 1 a des raisons de croire, sur cette base empirique, que le joueur 2 ira à droite à nouveau. Dans ce cas, il semble que le joueur 1 ne doive choisir son action que sur la base de sa croyance de premier ordre. S'il croit que le joueur 2 roulera à droite, sa meilleure réponse est de rouler à droite. Dans pareil cas, le joueur 1 n'a pas besoin de se représenter les croyances du joueur 2. Il lui suffit semble-t-il de se représenter l'action que choisira le joueur 2 pour déterminer son action.

Toutefois, le joueur 1 peut s'interroger sur ce qui justifierait l'action du joueur 2. Il pourrait se dire : « à mon avis, le joueur 2 ira plutôt à droite la fois prochaine, *car il est probable qu'il croit que j'irai à droite moi aussi* ». Dans ce cas, le joueur 1 se représente non plus seulement l'action probable du joueur 2, mais aussi les croyances du joueur 2 qui justifieraient l'action du joueur 2. Cette fois, le joueur 1 a des croyances sur les croyances du joueur 2, des croyances de second ordre.

Est-il utile d'aller plus loin ? La réponse est positive. Imaginons que le joueur 1 se dise : « j'ai des raisons spontanées de croire que le joueur 2 ira à droite. En particulier, le joueur 2 doit croire que j'irai à droite moi aussi. Mais n'allons pas si vite...*supposons que le joueur 2 puisse croire que je crois qu'il ira à gauche*. Dans ce cas, il devrait choisir de conduire à gauche. Et dans ce cas je devrais choisir de rouler à gauche ! ». Si donc le joueur 1 a des raisons de croire que le joueur 2 croit que le joueur 1 croit

que le joueur 2 ira à gauche, il doit choisir d'aller à gauche. Ici, le joueur 1 a des croyances de troisième ordre sur ce que fera le joueur 2. Naturellement, rien ne commande de s'arrêter à cette étape. Le joueur 1 pourrait se raviser une nouvelle fois : « imaginons que 2 croit que je crois que lui croit que j'irai à droite. Dans ce cas, 2 devrait plutôt choisir de rouler à droite. Et moi je devrais choisir de rouler à droite aussi ».

Chacune de ces étapes de raisonnement indique que, selon le niveau de croyance atteint, les agents sont susceptibles de se déterminer de différentes manières⁶. Plus encore, chaque niveau de croyance auquel s'arrête un agent est susceptible de lui suggérer une action plutôt qu'une autre. Une manière d'aboutir à de la coordination de façon sûre serait alors celle dans laquelle chacun croit que chacun roule à droite, chacun croit que chacun croit que chacun roule à droite, et ainsi de suite, *ad infinitum*. Une telle situation, dans laquelle les agents auraient toutes ces croyances ou attentes concordantes à tous les niveaux, suffit à réaliser ce que Lewis appelle *connaissance commune* (bien que ce ne soit pas la définition que retient Lewis de la connaissance commune, voir plus bas). En effet, la connaissance commune, comprise en ce sens, constitue alors une condition suffisante pour la coordination entre les agents.

Ce qui demeure une énigme cependant, c'est la question de savoir dans quelle mesure la connaissance commune ainsi entendue serait une condition nécessaire à la notion de convention. Après tout, comme l'indique le premier exemple de raisonnement donné plus haut, deux agents qui se contenteraient d'optimiser leurs actions simplement en répliquant leurs actions précédentes à l'équilibre, sans se soucier de ce que croit l'autre sur leurs croyances, ne pourraient-ils se maintenir à l'équilibre tout aussi bien ? Que faut-il de plus pour qu'un tel équilibre stable constitue une convention proprement dite ?

IV. Sélection et maintenance des équilibres

La question précédente peut être précisée de différentes manières. Skyrms (2004) distingue utilement deux formes du problème. Le premier est de déterminer si la connaissance commune est une condition nécessaire de l'émergence ou *sélection* d'un premier équilibre de coordination entre les agents. Le second est de déterminer si la connaissance commune est une

⁶ Voir notamment Clark et Marshall (1981), qui argumentent de la même façon que la connaissance commune de la référence des descriptions définies est nécessaire pour qu'il y ait communication réussie.

condition nécessaire de la perpétuation ou *maintenance* d'un équilibre une fois que les agents sont parvenus à cet équilibre. Lewis lui-même fait clairement la distinction entre ces deux formes du problème dans *Convention*. Il souligne que l'occurrence d'un premier équilibre de coordination peut résulter de différents processus : la *chance*, l'*accord* entre les agents, ou encore le *caractère saillant* de l'équilibre⁷.

Le fait d'atteindre un premier équilibre par chance, on le voit facilement, ne demande pas même que les agents aient des croyances sur ce que fera autrui. Il suffit que chacun choisisse son action en tirant à pile ou face, sans même songer aux raisons qui gouvernent le choix de l'autre. Le cas du choix d'une action sur la base d'un accord explicite, ou encore du fait du caractère saillant de telle ou telle action, en revanche, est plus clairement susceptible d'impliquer des croyances d'ordre supérieur. Dans le cas de l'accord explicite (cf. Lewis 1969, pp. 34 *sqq.*), nous sommes dans une situation où les agents peuvent se communiquer leurs intentions d'agir de telle ou telle manière. Ce serait un cas où le joueur 1 peut déclarer au joueur 2 son intention de rouler à droite, et idem pour le joueur 2. Comme le souligne Lewis, le fait de déclarer une intention ne suffit pas à assurer que l'action choisie sera conforme à l'intention. Mais le joueur 1, en déclarant son intention, crée une attente chez le joueur 2, et il s'attend à ce que le joueur 2 s'attende à ce qu'il agisse selon son intention, et ainsi de suite. La communication explicite, plus généralement, crée un système d'attentes mutuelles.

De la même façon, pour Lewis, le caractère saillant d'une action implique que chaque sujet aura tendance spontanément à la remarquer, mais aussi qu'il est capable de se représenter que l'autre aura cette même tendance à la remarquer. Imaginons le jeu de la conduite dans lequel le profil d'actions (G,G) donnerait pour gain (4,4), contre (3,3) pour le profil d'actions (D,D) (par exemple, rouler à gauche permet d'aller plus vite que rouler à droite). Le choix de l'action G est saillant ici car l'action est, en elle-même, plus avantageuse pour chaque agent que l'action D (dans les cas où personne n'arrive d'en face). Cela ne signifie pas que chacun choisira automatiquement l'action G, mais l'avantage qu'il y a à choisir G par rapport à D dans les cas où la route est libre peut suggérer à chaque joueur que l'autre peut s'attendre à ce qu'il choisisse G, et ainsi de suite.

⁷ Voir Lewis (1969), p. 36 : « Si vous et moi nous sommes rencontrés hier – par chance, par un accord, par saillance, ou quoi que ce soit – et qu'aujourd'hui nous découvrons que nous devons nous rencontrer de nouveau... » [*“ If you and I met yesterday – by luck, by agreement, by salience, or however – and today we find we must meet again....”*].

Comme le montrent ces exemples, la connaissance commune n'est donc pas une condition nécessaire de la sélection d'un premier équilibre, lequel peut se produire par chance. Pas même une croyance de second ordre n'est requise en principe dans les cas chanceux. En revanche, les cas d'accord et de saillance supposent déjà une théorie de l'esprit chez les agents, soit la capacité à se représenter les croyances d'autrui. Il est donc plausible que pour atteindre un premier équilibre, ces situations requièrent des croyances de niveau supérieur. Toutefois, la question de savoir si une hiérarchie infinie de croyances est requise reste en suspens.

Qu'en est-il alors de la maintenance des équilibres ? C'est pour ces cas-là que Lewis introduit la connaissance commune. A première vue, il semble qu'on puisse résumer sa position en disant : *la connaissance commune n'est pas une condition nécessaire de sélection d'un équilibre de coordination, en revanche, c'est une condition nécessaire de la maintenance de l'équilibre au statut d'une convention*. La raison intuitive est celle donnée dans la section précédente : si un agent peut douter des croyances de l'autre sur ses croyances, et ainsi de suite, à quelque niveau que ce soit, alors il est susceptible de choisir une action différente selon ce niveau. Sans un système de croyances concordantes à tous les niveaux, il semble qu'on ne puisse aboutir au choix univoque d'un profil d'actions.

Mais cette conclusion reste problématique. Skyrms par exemple met explicitement en doute l'idée selon laquelle la connaissance commune serait même condition nécessaire de la perpétuation des équilibres. Il écrit à propos des jeux de signal, vus comme des jeux de coordination spécifiques (Skyrms 2004, p. 55)⁸ :

« Pas si vite ! Avant de nous laisser trop troubler par le scepticisme philosophique, nous devrions nous souvenir que les bactéries ont développé et maintiennent des systèmes de signaux effectifs. Elles le font sans l'aide de la connaissance commune, de l'intentionnalité, ou du choix rationnel. Peut-être n'avons nous pas regardé le problème de la bonne façon. »

L'argument de Skyrms, pour l'essentiel, est celui esquissé à la fin de la section précédente. Pour qu'un équilibre de coordination se perpétue dans une situation d'interaction récurrente, il suffit sous certaines conditions que

⁸ « Wait a minute ! Before we let ourselves become too confused by philosophical skepticism, we should remember that bacteria have developed and maintain effective signaling systems. They do this without benefit of common knowledge, intentionality, rational choice. Perhaps we have not been looking at the problem in the right way »

les agents *répliquent* les actions à l'équilibre. Les modèles d'interaction répétée discutés par Skyrms sont plus élaborés, mais ils reposent fondamentalement sur cette intuition : que des agents exempts d'une théorie de l'esprit peuvent converger vers un équilibre stable à condition de répéter certaines actions en proportion du succès obtenu au cours de leurs interactions précédentes.

Ces considérations semblent extrêmement déflationnistes en regard de la théorie lewisienne du rôle de la connaissance commune, puisqu'elles mettent sur le même plan les conventions humaines et les régularités de comportement d'agents doués du minimum de rationalité concevable. Toutefois il nous semble hâtif de conclure que la notion de croyance d'ordre supérieur serait tout simplement superflue pour rendre compte de la notion de convention. On peut en effet opposer à Skyrms trois arguments de prudence. Le premier est un argument de *spécificité* : dès lors que l'on parle d'agents rationnels, capables d'avoir des croyances sur les croyances les uns des autres, quel rôle joue, pour ces agents, les croyances d'ordre supérieur dans la coordination ? Les modèles envisagés par Skyrms ne permettent pas de répondre à cette question. Le second argument est un argument qui concerne *l'efficacité éventuelle* des croyances d'ordre supérieur. Il est tout à fait possible que, pour qu'un équilibre se perpétue en une convention au sein d'une population d'agents rationnels, ces agents n'aient pas à utiliser beaucoup de théorie de l'esprit mais seulement des recettes inductives de plus bas niveau. Cependant, rien ne dit que la prise en compte des croyances d'autrui sur nos croyances ne soit pas requise pour aboutir de façon plus fiable à une convention que par l'emploi de stratégies qui seraient basées purement sur la réplication aveugle de certaines actions. Le troisième argument enfin concerne la *définition* même de la connaissance commune chez Lewis. Il y a dans les considérations de Lewis sur les croyances d'ordre supérieur une attitude elle-même déflationniste.

Au chapitre 1 de *Convention*, Lewis écrit en effet (p. 32) ⁹ :

Bien sûr, je n'imagine pas que quiconque résolve un problème de coordination en acquérant d'abord de quelque part une attente d'ordre 17

⁹ « *Of course, I do not imagine that anyone will solve a coordination problem by first acquiring a seventeenth-order expectation from somewhere and then sitting down to do his replications. For one thing, we rarely do have expectations of higher-order than, say, fourth. For another thing, any ordinary situation that could justify a high-order expectation would also justify low-order expectation directly, without recourse to nested replications* »

et en restant assis à faire ses répliques. D'une part, nous avons rarement des attentes d'ordre plus élevé que, disons, 4. D'autre part, une situation ordinaire qui pourrait justifier une attente d'ordre supérieur justifierait aussi une attente d'un ordre inférieur directement, sans le recours à des répliques enchâssées. »

Ce passage, peu remarqué des commentateurs de Lewis, indique deux choses. D'une part, s'agissant de la performance des agents, Lewis considère que les croyances d'ordre supérieur sont *de facto* limitées. S'agissant de leur compétence d'autre part, Lewis semble vouloir dire que, pour que des agents répliquent dans leurs croyances les croyances des agents sur leurs croyances afin de se coordonner, il n'est pas même nécessaire qu'ils se représentent explicitement cette hiérarchie de croyances. Afin de comprendre la théorie de Lewis, et pour voir en quel sens la connaissance commune pourrait être une condition nécessaire de coordination, il est temps d'entrer dans le détail de la définition par Lewis de la connaissance commune.

V. La connaissance commune

La *connaissance commune* d'une proposition q est à distinguer de la notion de *connaissance partagée*, qui sert à la définir. Une proposition q est dite connaissance partagée au sein d'un groupe si chacun dans le groupe sait que q . La définition la plus couramment utilisée de la connaissance commune (notamment telle que la présente informellement Aumann 1976) est la définition dite itérative qui implique non seulement que q soit connaissance partagée, mais encore que chacun sache que q est connaissance partagée, et ainsi de suite¹⁰. Autrement dit, la connaissance commune de q implique la vérité de q , implique que q soit connaissance partagée, implique qu'il soit connaissance partagée que q est connaissance partagée, et ainsi de suite à tous les niveaux.¹¹ La connaissance commune en ce sens est donc beaucoup plus exigeante que la connaissance partagée, puisque deux personnes par exemple peuvent chacune savoir quel temps il

¹⁰ La formalisation donnée par Aumann (1976) de la connaissance commune se fait en réalité en termes de partitions informationnelles et de plus petit grossissement commun aux partitions des agents, mais Aumann montre qu'elle est équivalente à la définition itérative.

¹¹ La définition de la *croyance* commune est la même, sauf que la croyance commune que p n'implique plus nécessairement la vérité de p . Notons que Lewis ne semble pas considérer que nécessairement, si un état de chose p indique à chaque agent que q , q est nécessaire une proposition vraie. En ce sens, la « connaissance » commune de Lewis peut être considérée comme de la croyance commune.

fait, sans savoir que chaque autre le sait.

La définition donnée par Lewis de la connaissance commune diffère toutefois de la définition itérative précédente. Chez Lewis une proposition q est dite connaissance commune au sein d'un groupe si et seulement s'il existe un état de choses p tel que :

- (1) chacun a des raisons de croire que p a lieu
- (2) p indique à chacun que chacun a des raisons de croire que p a lieu
- (3) p indique à chacun que q

Lewis appelle un tel état de choses p une *base* pour la connaissance commune que q . Le lien entre la définition par Lewis de la connaissance commune et la définition itérative n'est pas immédiat. Plus généralement, le lien entre la définition lewisienne de la connaissance commune et la notion de croyance d'ordre supérieur ne va pas de soi. La condition (1) en particulier n'implique que des croyances de premier niveau entre les membres du groupe. La condition (3) implique indirectement la notion de croyance, car Lewis entend que p indique à un agent que q si et seulement si les raisons de croire p pour l'agent sont par là même des raisons de croire q ¹². Mais là encore, (3) n'implique pas d'itération de croyance. La condition (2) en revanche a une structure récursive, et c'est notamment cette condition (2) qui permet de dériver la définition itérative de la connaissance commune¹³.

Cette dérivation toutefois n'est pas transparente à partir des hypothèses de Lewis, en partie du fait que Lewis ne propose pas de formaliser sa définition. Plusieurs tentatives de formalisation ont été faites ces dernières années, notamment par Cubitt et Sugden (2003), Sillari (2005), et Paternotte (2011). En nous inspirant de leur approche, une manière (néanmoins distincte) de formaliser les trois conditions de Lewis

¹² Voir Lewis (1969), p. 52-53 : « Qu'est-ce qu'indiquer ? Nous dirons que A indique à quelqu'un x que $_$ si et seulement si, si x avait une raison de croire que A se produisait, x aurait par là même des raisons de croire que $_$ » [« *What is indicating ? Let us say that A indicates to someone x that $_$ if and only if, if x had reason to believe that A held, x would thereby have reasons to believe that $_$.* »]

¹³ La condition (2) est parfois utilisée pour rapprocher la définition de Lewis d'une définition dite par point fixe de la connaissance commune, souvent jugée plus réaliste que la définition infinitaire. Voir Paternotte (2011). Sur la comparaison entre définition itérative, par point fixe, en en termes de situation partagée (notion proche de celle de base de connaissance commune chez Lewis), voir également Barwise (1988). Sur la définition dite par point fixe, cf. également Fagin et al. (1995).

serait la suivante, en représentant « i a des raisons de croire p » par l'opérateur $R_i p$:

- (1) pour tout i dans G : $R_i p$
- (2) pour tout i, j dans G : $R_i p \rightarrow R_i R_j p$
- (3) pour tout i, j dans G : $R_i p \rightarrow R_i q$

Ces trois conditions à elles seules ne suffisent pas à dériver la définition itérative de la connaissance commune. De (1) et (2) on peut dériver directement que $R_i R_j p$: chacun a des raisons de croire que chacun a des raisons de croire p . Pour dériver plus d'enchâssements, cependant, il faut des hypothèses plus fortes. Lewis précise toutefois (1969, p. 53)¹⁴:

« Les trois prémisses (1), (2), (3), moyennant des prémisses auxiliaires adéquates concernant notre rationalité, nos normes inductives et notre information d'arrière-plan, suffisent à justifier mes attentes d'ordre supérieur. »

Quelles sont ces prémisses auxiliaires auxquelles fait référence Lewis ? Pour Lewis, il s'agit en premier lieu de la capacité de chaque agent à se représenter que l'autre agent raisonne de la même façon que soi (ce que Lewis réfère au fait pour les agents de partager les mêmes normes inductives, et les mêmes informations d'arrière-plan). Cubitt et Sugden proposent un axiome spécifique pour capturer cette notion de mimétisme dans le raisonnement. Une possibilité explorée par Sillari (2005), qui nous semble plus simple et plus naturelle, est de supposer que l'opérateur R_i est un opérateur modal *normal*, qui satisfait la règle de nécessité et la clôture sous l'implication. Nous pouvons en particulier supposer que les schémas (2) et (3), qui sont les schémas proprement inférentiels (au contraire de (1)) obéissent à la règle dite de "nécessitation", c'est-à-dire que chaque agent a des raisons de croire que (2) et (3) valent pour lui et pour l'autre, et ainsi de suite à chaque niveau.

Formellement, la règle de nécessité (avec le principe de distributivité) suffit à dériver que tous les agents ont des raisons de croire que tous ont des raisons de croire que q , et ce pour un nombre arbitraire

¹⁴ « *The three main premises (1), (2), (3), together with suitable ancillary premises regarding our rationality, inductive standards, and background information, suffice to justify my higher-order expectations* ».

d'itérations de « tous ont des raisons de croire que ». ¹⁵ Cette hypothèse est très forte cependant, car elle revient en fait à supposer que les schémas (2) et (3) sont déjà connaissance commune. Conceptuellement, l'hypothèse peut se justifier néanmoins de la même manière que Lewis justifie informellement chaque étape de sa dérivation : chaque fois que *i* a des raisons de croire un certain énoncé de forme conditionnelle, alors s'il a des raisons de croire que *j* partage les mêmes normes d'inférence, il a des raisons de penser que *j* a ces mêmes raisons de croire, et ainsi de suite.

Tout cela ne suffit pourtant pas à dériver *exactement* la connaissance commune au sens de la définition itérative. Ce que nous pouvons dériver présentement est une hiérarchie infinie de *raisons de croire*. Pour en arriver à la hiérarchie de *croyances* proprement dit, Lewis note qu'il faut faire une hypothèse supplémentaire, à savoir que les agents sont ce que nous pouvons appeler « épistémiquement » rationnels, c'est-à-dire qu'ils croient effectivement ce pour quoi ils ont des raisons suffisantes, qu'ils croient que les autres sont aussi épistémiquement rationnels, que les autres croient que les autres le croient, et ainsi de suite ¹⁶. Comme l'écrit en effet Lewis (p. 57) ¹⁷,

« Une base de connaissance commune engendre des croyances d'ordre supérieur à l'aide d'attentes d'ordre supérieur préexistantes concernant la rationalité. »

En d'autres termes, pour que la hiérarchie de raisons de croire se traduise en une hiérarchie de croyances, il faut que la rationalité épistémique soit elle-même croyance commune, au sens itératif. Lewis lui-même insiste sur le fait que chaque étape supplémentaire au sein de la hiérarchie de croyances touchant la rationalité épistémique d'autrui constitue une

¹⁵ Par exemple: la règle de nécessité appliquée à (2) pour *j* implique $R_j(R_i p \rightarrow R_i R_j p)$ et par distributivité on obtient $R_j R_i p \rightarrow R_j R_i R_j p$. En y appliquant à nouveau la nécessité pour *i* on obtient: $R_i(R_j R_i p \rightarrow R_j R_i R_j p)$, et par distributivité : $R_i R_j R_i p \rightarrow R_i R_j R_i R_j p$, et ainsi de suite. De (1) et (2) on peut détacher chaque conséquent de cette chaîne d'implications. Semblablement, en appliquant la nécessité à (3), on obtient $R_j(R_i p \rightarrow R_i q)$, et par distributivité, $R_j R_i p \rightarrow R_j R_i q$, et en y appliquant la nécessité et la distributivité, $R_i R_j R_i p \rightarrow R_i R_j R_i q$, et ainsi pour chaque niveau.

¹⁶ Cf. Lewis (1969) p. 55 : « Quiconque a une raison de croire quelque chose en viendra à le croire, à condition qu'il ait un degré de rationalité suffisant » [“*Anyone who has reason to believe something will come to believe it, provided he has a sufficient degree of rationality*”]. Chaque degré de rationalité proprement dit semble donc correspondre au fait d'*actualiser* les étapes inférentielles de la dérivation donnée en termes de raison de croire (dans nos termes, au fait d'appliquer effectivement la règle de nécessité, la règle de distributivité, et la règle de détachement).

¹⁷ « *A basis for common knowledge generates higher-order expectations with the aid of pre-existing higher-order expectations of rationality* ».

hypothèse supplémentaire à propos de la rationalité des agents. C'est donc seulement dans l'idéal que ce qu'il nomme connaissance commune engendrera la hiérarchie complète de croyances. Selon lui, « le processus générateur s'arrête lorsque les prémisses auxiliaires cèdent » (1969, p. 56)¹⁸. Lewis considère en effet la dérivation de la définition itérative de la connaissance commune non comme un raisonnement actuel, mais comme une dérivation disponible en principe, pour des agents idéalement rationnels. Pour ce qui nous concerne, chaque niveau de croyance est atteint en principe à partir d'une nouvelle application de la règle de nécessité et d'une hypothèse supplémentaire à propos de la rationalité des agents, qu'on peut voir comme l'application par chaque agent de l'hypothèse que l'autre agent est au moins aussi rationnel que lui¹⁹. Mais de fait, on peut penser que des agents rationnels, étant limités dans leurs raisonnements, s'arrêteront au bout de quelques étapes de dérivation.

VI. Les fondements épistémiques de l'équilibre de Nash

Les considérations de la section précédente nous éclairent sur la notion lewisienne de connaissance commune. Comme nous l'avons souligné, Lewis ne prend pas la définition itérative de la connaissance commune comme point de départ. Mais sa définition de la notion d'événement base de la connaissance commune est censée, pour des agents qui seraient parfaitement rationnels, aboutir à cette définition²⁰. De fait,

¹⁸ « *The generating process stops when the ancillary premisses give out* ».

¹⁹ Notons que Lewis conçoit le niveau des attentes de rationalité comme étant lui-même une fonction de la situation dans laquelle se trouvent les agents. A propos des attentes de niveau supérieur sur la rationalité, il écrit : « Celles-ci peuvent-elles elles-mêmes être engendrées à partir d'une base de connaissance commune ? » [« *Can these themselves be generated by some basis for common knowledge* »]. Il répond : « Oui, parce que les attentes d'ordre supérieur en matière de rationalité qui sont requises pour engendrer une attente d'ordre n sont elles-mêmes d'ordre moins que n . Ce qui interrompt l'engendrement des attentes d'ordre supérieur, c'est la quantité limitée de rationalité indiquée par une base – non pas la difficulté à engendrer des attentes d'ordre supérieur d'autant de rationalité qu'en indique une base » [« *Yes, because all the higher-order expectations of rationality needed to generate an n -th-order expectation are themselves of less than n -th-order. What cuts off the generation of higher-order expectations is the limited amount of rationality indicated by any basis – not any difficulty in generating higher-order expectations of as much rationality as is indicated by any basis.* »] Cette remarque suggère que le niveau auquel la règle de nécessité s'appliquera, dans notre reconstruction (soit l'ordre auquel les agents répliquent (2) et (3)), peut lui-même dépendre de la nature de la base en question

²⁰ Le fait que Lewis accepte la définition itérative de la connaissance commune est confirmé par les autres textes de Lewis où celui y fait référence. Cf. en particulier Lewis (1973, p.165), et aussi Lewis (1972, pp. 257-258), où Lewis écrit : « Collectez toutes les platitudes auxquelles vous pouvez songer concernant les relations causales entre

Lewis accepte la définition itérative, mais considère que ce qui compte, pour qu'il y ait convention, c'est d'avoir de la connaissance commune *potentielle*, non pas nécessairement *actuelle*. Une base de connaissance commune est un événement qui donne de la connaissance commune potentielle, dans la limite des capacités des agents à engendrer des croyances d'ordre supérieur sur cette base²¹. Pour Lewis, ainsi, l'accord entre agents, le caractère saillant d'une action, ou encore le précédent lié à la réalisation d'un équilibre antérieur, sont des bases de connaissance commune. Lewis donne notamment l'exemple de la conduite à droite comme un tel exemple²²:

« Par exemple, les conducteurs aux Etats Unis ont jusqu'ici conduit à droite. Nous avons tous des raisons de croire qu'il en va ainsi. Et le fait qu'il en soit ainsi nous indique à tous que nous tous avons des raisons de croire que les conducteurs aux Etats-Unis ont jusque là conduit à droite et aussi que les conducteurs auront tendance à conduire à droite dans le futur. »

Ces considérations de Lewis sont psychologiquement vraisemblables, mais elles ne nous permettent pas de répondre en toute généralité à la question que nous posions dans les sections précédentes, à savoir : les croyances de niveau supérieures sont-elles nécessaires en principe pour sélectionner ou pour maintenir un équilibre de coordination ? Si oui, faut-il de la connaissance commune potentielle, c'est-à-dire faut-il que les agents puissent en principe avoir des croyances concordantes à tous les niveaux ?

Pour continuer d'éclairer cette question, nous proposons de confronter notre perspective aux principaux résultats dits de caractérisation épistémique de l'équilibre de Nash en théorie des jeux. Avant de présenter ces résultats, et d'évaluer leur pertinence pour la compréhension de la notion lewisienne de convention, il vaut d'abord la peine d'expliquer

états mentaux, stimuli sensoriels et réponses motrices. (...) Ne comptez que les platitudes qui sont connaissance commune parmi nous – chacun les sait, chacun sait que chacun les sait, et ainsi de suite. » [« *Collect all the platitudes you can think of regarding the causal relations of mental states, sensory stimuli, and motor responses. (...) Include only platitudes which are common knowledge among us – everyone knows them, everyone knows that everyone knows them, and so on.* »].

²¹ Voir Lewis (1973, p. 165), où Lewis écrit : « La connaissance mentionnée ici peut être purement potentielle : de la connaissance qui serait disponible si l'on s'efforçait d'y penser suffisamment ». [« *The knowledge mentioned here may be merely potential : knowledge that would be if one bothered to think hard enough.* »]

²² « *For example, drivers in the Unites States have hiterto driven on the right. All of us have reason to believe that this is so. And the fact that this is so indicates to all of us that all of us have reason to believe that drivers in the United States will tend to drive on the right henceforth.* ».

brèvement en quoi consistent de telles caractérisations.

Traditionnellement, les concepts de solution en théorie des jeux sont présentés comme des hypothèses, normatives ou descriptives, sur les résultats possibles de l'interaction. L'équilibre stratégique, que nous avons examiné jusqu'ici, n'est pas le seul concept de solution étudié en théorie des jeux, mais il illustre bien cette conception²³. Par exemple, comme nous l'avons mentionné plus haut, le jeu correspondant à la matrice en page 3 contient deux équilibres²⁴, (G, G) et (D, D), où les joueurs conduisent du même côté de la route. Selon l'interprétation classique des concepts de solutions, ces deux équilibres sont ce que des agents devraient rationnellement jouer (interprétation normative) ou ce qu'ils joueront effectivement (interprétation descriptive).

Dans quelles conditions des agents éventuellement rationnels joueront-ils un certain concept de solution ? En termes plus spécifiques, existe-t-il des conditions sous lesquelles nous pouvons légitimement supposer que les actions des agents impliqués dans une interaction stratégique donnée seront à l'équilibre ?

Une réponse que nous pouvons appeler « pré-épistémique » est d'en appeler aux principes de rationalité développés en théorie de la décision individuelle. Dans cette perspective, l'idée fondamentale est que la théorie des jeux, et en particulier les concepts de solution, supposent la « connaissance commune de la rationalité »²⁵. Vue en termes classiques, la rationalité signifie alors simplement la maximisation de l'utilité espérée,

²³ Un autre exemple de concept de solution est notamment la *dominance stricte itérée*. L'idée est que, dans un jeu, un agent rationnel peut éliminer celles de ses actions qui sont strictement moins profitables que d'autres en toute circonstance. L'un des résultats les plus cités sur la caractérisation épistémique des concepts de solution est celui d'après lequel la connaissance commune de la rationalité implique l'élimination itérée des stratégies strictement dominées. Dans *Convention*, Lewis mentionne ce concept de solution (1969, p. 19), et il est remarquable qu'il signale déjà le lien avec la notion de connaissance commune. Il écrit ainsi : « By successive deletions of dominated actions, the game is transformed into a situation that is patently trivial because Row-chooser and Column-chooser each have only one available action. *The outcome is determined by the fact that everybody ignores dominated actions, everybody expects everybody else to ignore dominated actions, and so on.* » (souligné par nous). Le point important est que dans les jeux de coordination qui intéressent Lewis, qui comportent des équilibres multiples, il n'existe pas de stratégie strictement dominante. Le concept de dominance stricte itérée est ainsi trop restrictif pour définir les conventions, ce qui explique que Lewis lui préfère le concept d'équilibre de Nash. Pour une comparaison approfondie des fondements épistémiques de l'équilibre de Nash par opposition à la dominance stricte, nous renvoyons à Perea (2012).

²⁴ Ce sont des équilibres en stratégies dites « pures ». Ce jeu a aussi un équilibre en stratégies dites mixtes, c'est-à-dire où chaque joueur joue chacune une ces stratégie avec une probabilité de 0.5.

²⁵ Voir par exemple Meyerson (1997).

dans la lignée des travaux de Ramsey, de Finetti, von Neumann et Morgenstern, ou encore Savage sur le choix rationnel. Cette réponse est « pré-épistémique » au sens où la notion de connaissance commune n'y est pas formalisée. L'accent est mis avant tout sur les concepts de solution. La notion de croyance elle est largement absente des modèles classiques en théorie des jeux. L'équilibre de Nash, par exemple, n'est défini qu'en termes de préférences, compte-tenu de ce que les agents choisissent. Pour revenir à notre exemple, à l'équilibre, étant donné qu'un des joueurs roule d'un certain côté de la route, l'autre préfère rouler de ce même côté, et vice-versa. Cette description de l'équilibre fait appel aux préférences des joueurs, mais elle ne fait pas explicitement référence à leurs croyances respectives à propos des actions des autres, dimension pourtant cruciale pour procéder au calcul de l'utilité espérée.

La théorie épistémique des jeux, développée notamment par Aumann et ses collaborateurs, peut être vue comme une tentative pour rétablir, cette fois d'une manière rigoureuse, ce lien entre la théorie de la rationalité en situation d'interaction et les concepts et outils développés en théorie de la décision. Partant de l'idée que la prise de décision en situation d'interaction n'est pas essentiellement différente de la prise de décision en situation d'incertitude, la théorie épistémique des jeux étudie les « conséquences comportementales » des hypothèses qu'on peut faire sur la rationalité et les croyances mutuelles des joueurs en situation d'interaction. Cette approche de la théorie des jeux propose d'identifier quelles actions seront jouées ou exclues par des agents choisissant selon une règle de décision particulière, étant données certaines croyances à propos des autres agents impliqués dans la même situation (comme par exemple la croyance qu'ils suivent eux aussi la même règle de décision).

Naturellement, les concepts de solution figurent en tête des « conséquences comportementales » étudiées en théorie épistémique des jeux. Dans ce cas particulier on parle de caractérisations épistémiques de ces concepts. Le but est d'identifier des conditions nécessaires ou suffisantes sur la rationalité et les croyances des agents pour que les agents jouent selon un concept de solution donné. Le cas qui nous intéresse concerne les résultats de caractérisation épistémique de l'équilibre de Nash, en particulier parce que le rôle joué par le concept de connaissance commune y est plus complexe qu'on ne pourrait s'y attendre à première vue.

VII. Les résultats d'Aumann et Brandenburger

Dans ce qui suit nous nous penchons sur les deux caractérisations les plus connues de l'équilibre stratégique, dues à Aumann et Brandenburger (1995). La première, la plus simple, s'applique seulement aux jeux à deux joueurs. La seconde, plus générale mais par là même plus complexe, s'applique à tout jeu ayant un nombre fini de joueurs.

Tout au long de cette section, nous ferons un certain nombre d'hypothèses, sur lesquelles nous reviendrons dans la section suivante. Nous supposerons tout d'abord que les agents sont épistémiquement rationnels au sens discuté plus haut, c'est-à-dire qu'ils croient ce pour quoi ils ont des raisons suffisantes. Fidèles à la démarche de la théorie épistémique des jeux, nous supposerons de plus que chaque agent est un agent « bayésien », au sens où ses croyances partielles peuvent être représentées par une fonction de probabilité sur les états de choses possibles qui déterminent les conséquences de ses actions, en l'occurrence les actions (et croyances partielles) des autres joueurs. De ce fait nous endossons certaines hypothèses supplémentaires concernant la rationalité épistémique des agents, à savoir que leurs croyances satisfont les critères de cohérence (probabilités) classiques. Encore une fois, il est crucial de distinguer cette forme de rationalité, proprement épistémique, de la notion plus pratique de rationalité associée à la maximisation de l'utilité espérée. Un agent peut très bien avoir des croyances partielles tout à fait cohérentes et néanmoins choisir une action « irrationnelle » au sens où elle ne maximiserait pas l'utilité espérée.

Dans cette représentation probabiliste, nous identifierons la croyance que P avec le fait pour cet agent d'être certain que P, au sens où il assigne une probabilité de 1 à cette proposition (ensemble d'états possibles du monde). Cette identification, courante en théorie épistémique des jeux, est éminemment critiquable mais elle joue ici un rôle simplificateur. Nous indiquerons au passage lorsque cette supposition peut être levée sans compromettre les résultats présentés.

Enfin, dans tous ces résultats, nous utiliserons la notion itérative de la connaissance commune, dans l'une ou l'autre de ses versions formalisées depuis Aumann, plutôt que la notion plus complexe de Lewis. La raison en est que, premièrement, c'est la notion d'Aumann qui fut utilisée pour formuler ces résultats et que, deuxièmement, comme nous l'avons déjà souligné, la définition itérative est en réalité acceptée par Lewis pour les cas idéaux où la rationalité épistémique est connaissance commune.

VII. 1. Les conditions de l'équilibre dans les jeux à deux joueurs

Revenons maintenant à la matrice du jeu de la conduite (cf. Section 2), et supposons que le joueur 1 a de bonnes raisons de croire que 2 conduira à gauche. Comme nous l'avons mentionné plus haut, ceci peut provenir du fait que 1 et 2 se sont souvent rencontrés sur cette route et que 2 a toujours conduit à gauche, ou simplement que 1 est au fait de l'existence préalable d'une convention gérant la conduite sur cette route. Soit p et $(1 - p)$ les valeurs de probabilité que 1 assigne respectivement aux propositions G_2 , que 2 roulera à gauche, et D_2 , que 2 roulera à droite. Dénotons de plus $u_i(a_1, a_2)$ la valeur d'utilité que le joueur i obtient lorsque 1 joue une certaine action a_1 et 2 joue l'action a_2 . Par exemple, $u_1(G, G)$ est simplement 1, c'est-à-dire la valeur d'utilité que 1 attache à l'issue où les deux joueurs conduisent à gauche (la case supérieure gauche de la matrice en début de texte). L'utilité espérée de chacune des actions de 1 s'exprime comme la somme pondérée des valeurs d'utilité attachées aux conséquences de ces actions.

$$UE_1(G) = p \cdot u_1(G, G) + (1 - p) \cdot u_1(G, D)$$

$$UE_1(D) = p \cdot u_1(D, G) + (1 - p) \cdot u_1(D, D)$$

En fixant la valeur de p à 1, représentant le fait que 1 croit que 2 conduira à gauche, et en remplaçant les valeurs d'utilité, on obtient :

$$UE_1(G) = 1(1) + 0(0) = 1$$

$$UE_1(D) = 1(0) + 0(1) = 0$$

Compte-tenu du fait que 1 croit que 2 conduira à gauche, conduire lui aussi à gauche est la seule action qui maximise l'utilité espérée²⁶. La situation est entièrement symétrique dans le cas où le joueur 1 croit que 2 conduira à droite, et de même, vice-versa, pour le joueur 2.

Dans ces deux calculs, nous avons utilisé la notion de croyance. Or, si nous supposons de plus que les croyances des agents sont vraies, c'est-à-dire que si 1 croit que 2 conduit à gauche, 2 conduit effectivement de ce côté de la route, alors leurs actions seront à l'équilibre. En d'autres termes,

²⁶ Notons que dans cet exemple, nous aurions pu choisir n'importe quelle valeur de p supérieure à 0.5 pour représenter la croyance du joueur 1.

si chacun des joueurs a une croyance correcte concernant ce que fait l'autre joueur, et que chacun choisit une action qui maximise l'utilité espérée, alors ils joueront l'un des équilibres, c'est-à-dire qu'ils coordonneront leur choix soit sur (G, G) soit sur (D, D).

Cette observation se généralise à n'importe quel jeu à deux joueurs : *si chaque joueur choisit une action qui maximise l'utilité espérée, dans un contexte où cet agent sait quelle action l'autre choisit, alors les choix des deux joueurs sont à l'équilibre.* En d'autres termes, la rationalité (comme maximisation de l'utilité espérée) et la connaissance partagée des actions choisies impliquent un jeu à l'équilibre. Ce résultat, dû à Aumann et Brandenburger (1995), est la première caractérisation épistémique de l'équilibre stratégique qui nous intéresse. Avant d'en évaluer la pertinence pour la question de l'émergence comme du maintien des conventions au sens lewisien, quelques remarques s'imposent.

Premièrement, la notion de véracité de la croyance est centrale pour ce résultat. Dans l'exemple ci-dessus, si les deux joueurs choisissent l'action qui maximise l'utilité espérée compte tenu d'une croyance erronée, le résultat sera un profil hors équilibre, soit (G, D) ou (D, G). Il faut noter cependant que le résultat ci-dessus est général au sens où il s'applique à tout jeu à deux joueurs représenté sous forme matricielle, y compris pour les jeux qui ne sont pas des jeux de coordination au sens de Lewis. Le cas des jeux de coordination permet de renforcer le résultat. *Dans un jeu de coordination où pour chaque action d'un joueur il y a une réponse optimale unique à l'action de l'autre, il est suffisant, pour garantir l'équilibre, qu'un seul des deux joueurs ait une croyance vraie à propos de ce que fait l'autre.*

Deuxièmement, il est remarquable que ce résultat, dans sa forme générale comme dans sa forme s'appliquant spécifiquement aux jeux de coordination, ne fasse pas appel, du moins de manière explicite, à la notion de connaissance commune, aussi bien en ce qui concerne les choix que la rationalité des joueurs. Pour nous en convaincre, supposons, par exemple, que le joueur 1 ait la croyance erronée que 2 croit que 1 jouera G, mais qu'il joue néanmoins D. Ce choix est pour lui irrationnel, compte-tenu de cette croyance. Toutefois, si 2 a une croyance correcte à propos du choix de 1, il jouera lui aussi D. Le résultat sera donc à l'équilibre, et ce malgré le fait que, premièrement, 1 ait des croyances erronées à propos des croyances de 2 et que, deuxièmement, 1 n'ait pas choisi rationnellement. Ni leur choix respectif ni la rationalité des joueurs ne sont connaissance commune dans ce cas.

La notion de connaissance commune tient cependant un rôle sous-

jacent dans ce résultat. « L'espace d'incertitude », c'est-à-dire l'ensemble des états de choses possibles, incluant les faits à propos des croyances possibles des joueurs, est « connaissance commune ». « Connaissance commune » est ici entre guillemets parce que l'espace d'incertitude lui-même n'est pas une proposition ou un événement à propos duquel les agents peuvent former des croyances explicites dans les modèles qui sont sous-entendus ici. Cette notion de connaissance commune de l'espace d'incertitude est une donnée externe au modèle. Reste néanmoins que tous les agents, dans ces résultats, raisonnent à partir du même modèle²⁷.

Cela n'implique pas que les croyances réelles des agents soient connaissance commune, mais plutôt que l'ensemble des croyances possibles le soit. Ce présupposé supplémentaire ne joue pas un rôle important dans le résultat pour les jeux à deux joueurs que nous venons de présenter, mais il importe pour sa généralisation à un nombre arbitraire fini de joueurs, sur laquelle nous nous pencherons sous peu.

Alors que l'espace d'incertitude est connaissance commune dans ce résultat, la structure du jeu elle-même, à savoir les actions possibles pour chaque joueur, et leurs préférences, ne l'est pas nécessairement. Dans l'exemple que nous venons de discuter, nous avons supposé implicitement que 1 sait que les préférences de 2 sont identiques aux siennes, et vice-versa. Mais cette supposition n'est pas nécessaire. Si 1 sait ce que fait 2, et que 1 est rationnel, il n'a besoin d'aucune information supplémentaire, par exemple à propos de ce que 2 préfère, pour calculer quelle action maximise son utilité espérée. Nous n'avons en fait pas utilisé cette information dans le calcul plus haut. Il est intéressant ici de noter la différence avec la notion lewisienne de convention, qui suppose qu'il soit connaissance commune que « chacun préfère [coordonner son choix sur ceux des autres] à la condition que les autres le fassent ». En termes plus techniques, ce résultat de Aumann et Brandenburger tient aussi pour les jeux dits à information incomplète, où il peut y avoir de l'incertitude sur la structure du jeu elle-même, alors que la notion de convention chez Lewis semble s'appliquer seulement aux jeux à information complète. Nous reviendrons sur ce point, car faire, avec Lewis, la supposition supplémentaire que les préférences des agents sont connaissance partagée a des conséquences inattendues lorsqu'on examine les conditions épistémiques pour l'équilibre dans les jeux à plus de deux joueurs.

²⁷ Modèle qui, par ailleurs, peut être très complexe, et donc représenter un haut degré d'incertitude réciproque à propos de ce que chaque agent considère possible.

VII. 2. Le cas de l'équilibre à n joueurs

Rationalité et connaissance partagée des actions choisies ne sont pas suffisantes pour assurer un jeu à l'équilibre lorsqu'il y a plus de deux joueurs. Le résultat général d'Aumann et Brandenburger, pour un nombre arbitraire fini de joueurs, est le suivant. *Supposons que dans une situation particulière de jeu, les agents aient ce qu'on appelle des croyances antérieures communes. Supposons, de plus, que les préférences des agents, ainsi que le fait qu'ils sont rationnels, soient connaissance partagée. Si, dans une telle situation, les « conjectures » de chacun à propos de ce que font les autres sont connaissance commune, alors ces conjectures forment un équilibre.*

Deux notions sont centrales pour comprendre ce résultat, celle de croyance antérieure commune (par quoi nous traduisons *common prior belief*) et celle de conjecture²⁸. Commençons par la notion de *conjecture*. On entend par là une distribution de probabilité, pour chaque joueur, sur les actions des autres. Le cas de la croyance (correcte) avec probabilité 1 qui vient d'être évoqué est un cas particulier de conjecture où chaque agent est certain que l'autre jouera une stratégie pure particulière. En général, cependant, ces conjectures assigneront une valeur de probabilité non-extrême, strictement entre 0 et 1, à l'ensemble de stratégies pures des autres joueurs.

Le théorème est remarquable en ceci qu'il traite une notion proprement épistémique et subjective, à savoir celle de conjecture, comme constitutive de l'équilibre. On passe d'un équilibre d'actions à un équilibre de croyances. Ce glissement est intentionnel. Il repose sur ce qu'on appelle l'interprétation épistémique de la notion de stratégie mixte. Rappelons qu'une stratégie mixte est une distribution de probabilité sur un ensemble de stratégies pures. Dans notre exemple, la stratégie pour le joueur 1 consistant à jouer D et G avec la même probabilité (50%) est une stratégie dite mixte. Dans ce cas il est facile d'interpréter ce « choix » en termes objectifs. Le joueur 1 tire à pile ou face et laisse le résultat décider de l'action qu'il entreprendra²⁹.

²⁸ La notion de *croyance antérieure commune* ne doit pas être confondue avec la notion de *croyance commune* au sens de Lewis. « Commune » dans ce cas signifie simplement que les probabilités antérieures sont les mêmes pour tous les joueurs.

²⁹ Notons que cette interprétation « naturelle » ne justifie pas le choix d'une stratégie mixte plutôt que pure. Cette question a fait l'objet de nombreux débats en théorie des jeux et en philosophie. Nous la laisserons cependant de côté parce que, comme nous le

Le concept de stratégie mixte admet cependant n'importe quelle distribution de probabilité sur un ensemble donné de stratégies pures, et dans bien des cas il est peu réaliste de supposer que les agents ont à leur disposition le mécanisme aléatoire précis leur permettant de « jouer » une telle stratégie. L'interprétation épistémique évite ce problème en concevant une stratégie mixte pour un joueur non pas comme le résultat d'un processus aléatoire concret, mais comme représentant les croyances des autres joueurs à propos des actions de ce joueur. Selon cette interprétation, la stratégie mixte pour 1 consistant à conduire à gauche avec une probabilité de 50% est comprise comme reflétant les croyances de 2 à propos de ce que fera 1. En l'occurrence 2 considère également probable que 1 conduise à droite ou à gauche.

Dans l'interprétation épistémique des stratégies mixtes, chaque joueur ne choisit que parmi son ensemble de stratégies pures. Retournons à notre exemple, et considérons l'équilibre en stratégies mixtes où chaque joueur conduit de chaque côté de la route avec la même probabilité, 50%. Selon l'interprétation épistémique des stratégies mixtes, ces valeurs de probabilité sont vues comme les croyances de chaque joueur à propos de ce que fera l'autre. Comme nous venons de le mentionner, la stratégie mixte pour 1 consistant à conduire à gauche avec une probabilité de 50% est donc comprise comme reflétant les croyances de 2 à propos de ce que fera 1. Prenons maintenant la perspective de 2. Compte tenu de cette croyance, 2 peut rationnellement jouer G ou D (et vice-versa pour 1) :

$$UE_2(G) = 1/2(1) + 1/2(0) = 1/2$$

$$UE_2(D) = 1/2(0) + 1/2(1) = 1/2$$

En particulier, ce système de croyances à l'équilibre sanctionne aussi les points de non-coordination (D, G) et (G, D). Ce résultat est à première vue contre-intuitif. Il reflète cependant l'idée que sans information particulière sur ce que fera l'autre, les joueurs n'ont aucune raison de choisir une action plutôt qu'une autre. Compte-tenu de ce système de croyances, aucun joueur n'a intérêt à choisir autrement. C'est le système de conjectures dans sa totalité qui rationalise les actions à l'équilibre.

Avant de discuter plus en détail le rôle de la connaissance commune dans ce résultat, il est important de s'attarder sur les autres notions qui sont

verrons, sous l'interprétation épistémique utilisée dans la caractérisation épistémique de l'équilibre cette question est évitée. Ajoutons que Lewis, dans *Convention*, ne semble pas accorder de place à la notion de stratégie mixte.

la clef-de-voûte de ce théorème. Le théorème suppose la connaissance partagée de deux paramètres, à savoir la structure de préférences des joueurs et leur rationalité. Nous avons vu plus haut qu'aucun présupposé de la sorte n'était nécessaire dans les jeux à deux joueurs. On fait donc un pas, en supposant ce niveau de connaissance partagée, vers la supposition faite par Lewis que la structure de préférences dans le jeu est elle-même connaissance commune³⁰. Il en est de même, de prime abord, pour la connaissance commune de la rationalité.

Fait étonnant, cependant, si comme Lewis on suppose que la structure du jeu, et en particulier des préférences, est connaissance commune, alors les conditions citées dans le théorème ci-dessus impliquent ensemble que la rationalité des agents sera elle-aussi connaissance commune³¹. En d'autres termes, en ajoutant aux prémisses du résultat ci-dessus que les préférences de chacun sont non seulement connaissance partagée mais encore connaissance commune, on obtient non seulement que les conjectures des agents sont à l'équilibre, mais que les agents ont aussi connaissance commune de leur rationalité.

Cela étant dit, il est important de noter que ceci ne revient pas à affirmer que la connaissance commune de la rationalité est à elle seule suffisante pour que les agents jouent à l'équilibre. Le résultat est *strict* («*tight*») au sens où il tombe dès qu'une des conditions cités plus haut n'est pas atteinte. En particulier, pour n'importe quel niveau dans la hiérarchie infinie de connaissance partagée des conjectures, il existe des situations où les conjectures ne sont pas connaissance partagée jusqu'à ce niveau et où les agents ne jouent pas à l'équilibre, et ce même s'il est connaissance commune que tous sont rationnels et qu'il n'y a aucune incertitude sur les préférences. Il est important de noter que le caractère strict du résultat ne rend pas ces conditions nécessaires à l'atteinte de l'équilibre. Si l'une de ces conditions tombe dans une situation particulière, il ne s'ensuit pas nécessairement que les conjectures des agents ne sont pas à l'équilibre, ce qui devrait être le cas si les dites conditions étaient aussi nécessaires. Ces conditions sont suffisantes à l'équilibre dans le sens (strict) suivant : si l'une d'entre elles ne tient pas alors il est possible (mais il ne s'ensuit pas) que les conjectures des agents ne soient pas à l'équilibre.

L'autre pilier fondamental dans ce résultat est la notion de *croyances*

³⁰ Comme nous l'avons noté en dans la Section 2, la définition finale de la notion de convention proposée par Lewis exige que les conditions 1 à 3 soient connaissance commune. La condition 3 décrit la structure de préférences constitutive d'un jeu de coordination.

³¹ Cf. Polack (1999).

antérieures communes (« *common prior beliefs* »). Rappelons que nous supposons ici que les agents sont bayésiens au sens où leurs croyances sont représentées de manière probabiliste. La notion de croyance antérieure pour un agent réfère aux attentes *a priori* de cet agent, c'est-à-dire les attentes qu'il possédait avant de recevoir quelque information particulière que ce soit à propos d'une situation donnée. Ces attentes peuvent être totalement neutres, donnant à chaque état de choses possible la même valeur de probabilité, mais elles peuvent tout aussi bien être inclinées vers certains états de choses plutôt que d'autre, reflétant des faits aussi variés et larges que le contexte socio-historique dans lequel cet agent est situé. Ce sont ces croyances antérieures qui formeront la base à partir de laquelle cet agent formera ses croyances postérieures par conditionnalisation lorsqu'il recevra de l'information plus spécifique à propos d'une situation donnée.

La notion de croyance antérieure commune correspond à l'idée que toute différence d'opinion, c'est-à-dire toute différence dans les croyances postérieures entre les agents au sein d'un groupe donné, doit être vue comme relevant de différences dans l'information privée au moyen de laquelle chaque agent a formé ses croyances postérieures³². En d'autres termes, des agents bayésiens ayant des croyances antérieures communes devraient tirer les mêmes conclusions s'ils sont pourvus d'information identique. Ces agents ont, *a priori*, les mêmes attentes, et mettent à jour leurs croyances de la même façon. Notons que ce présupposé n'implique pas que les agents auront des croyances postérieures identiques. Il est tout à fait possible que des agents aient des croyances antérieures semblables mais que, sur la base d'informations privées différentes, ils forment des croyances postérieures diamétralement opposées sur un sujet donné.

Dans quels contextes est-il justifié de supposer que les agents ont des croyances antérieures communes ? La question est délicate, et il serait beaucoup trop long ici de couvrir les discussions détaillées dont elle a fait l'objet en philosophie et en théorie économique. Contentons-nous de mentionner deux points. Tout d'abord, la notion de croyance antérieure commune est souvent vue comme un présupposé méthodologique en économie, destiné à prévenir une certaine forme de trivialisation. Sans ce présupposé, à peu près n'importe quel comportement peut être rationalisé sur la base de croyances préalables différentes, souvent calculées de manière *ad hoc*³³. Ensuite, il existe aujourd'hui plusieurs résultats mathématiques montrant que sous certaines conditions relativement plausibles, les

³² Voir Harsanyi (1967-68), Aumann (1987) et Bonnano et Nehring (1999).

³³ Voir S. Morris (1995) et G. Bonanno et K. Nehring (1999).

croyances postérieures des agents dans un groupe peuvent toujours être vues comme provenant de croyances antérieures communes³⁴. Le présupposé que les agents ont des croyances antérieures communes mérite certes une analyse plus détaillée, mais pour la suite nous procéderons en supposant qu'il peut être justifié de façon indépendante.

Dans la caractérisation épistémique de l'équilibre, la notion de croyance antérieure commune joue un rôle fondamental en combinaison avec la supposition que les conjectures des agents sont connaissance commune. De prime abord, les conjectures de chacun à propos du choix d'un agent particulier peuvent différer grandement. Il n'y a aucune raison, du moins *a priori*, de supposer que tous les agents auront les mêmes attentes à propos de ce que feront les autres. Cela pose évidemment un problème d'indétermination pour le résultat que nous discutons ici. Si ces conjectures sont effectivement différentes, il n'y a rien de tel, sauf dans les jeux à deux joueurs, que « la » conjecture des autres à propos de ce que fera un agent donné, laquelle est au fondement de l'interprétation épistémique des stratégies mixtes que nous venons de présenter.

La notion de croyance antérieure commune permet d'éviter ce problème d'indétermination. Par le « théorème d'accord » établi par Aumann (1976), si les agents d'un groupe donné ont des croyances antérieures communes, et que leur croyances postérieures sont connaissance commune, alors ces croyances postérieures sont identiques³⁵. Dans le contexte qui nous occupe, les croyances postérieures sont les croyances des agents à propos de ce que feront les autres, les conjectures. En supposant que ces conjectures sont connaissance commune, et que les agents ont formé ces conjectures sur la base de croyances antérieures communes, on obtient par le théorème d'accord que ces conjectures doivent être les mêmes au sein du groupe.

Deux points importants concernant cette application du théorème d'accord à la caractérisation des équilibres pour les jeux à plus de deux joueurs. Premièrement, répétons que le théorème d'accord n'est appliqué ici qu'aux croyances des agents à propos des actions des autres. Rien n'est dit sur les autres croyances des agents, croyances qui elles peuvent varier grandement. Rappelons en effet de la supposition que les agents ont des croyances antérieures communes ne contraint que peu ou pas les croyances postérieures auxquelles chacun peut arriver après avoir reçu de

³⁴ Voir J. Halpern (2002), et A. Heifetz (2006)

³⁵ Pour plus de références sur le théorème d'accord, nous renvoyons à G. Bonanno et K. Nehring (1997).

l'information privée. Ensuite, cette application du théorème d'accord place le fardeau de la preuve philosophique de manière encore plus nette sur la justification de la notion de croyances antérieures communes. Nous avons esquissé plus haut quelques raisons générales militant en faveur de cette supposition, mais nous y reviendrons dans la discussion qui suit.

VII. 3. Connaissance commune et équilibre : résumé

La relation entre équilibre stratégique et connaissance commune est donc complexe et variée. Résumons-en les faits saillants avant de discuter de leur pertinence pour notre compréhension de la notion de convention. Dans un premier temps, pour les jeux (de coordination) à deux joueurs, ni la connaissance commune des actions jouées ni celle de la rationalité ne sont nécessaires à l'équilibre. Les notions-clés y sont celles de rationalité et de connaissance partagée des choix. La connaissance commune joue un rôle beaucoup plus marqué en revanche dans les jeux à plus de deux joueurs. La caractérisation de l'équilibre y fait appel à la connaissance commune des conjectures, et ce sont ces conjectures elles-mêmes, vues comme des croyances mutuelles à propos de ce que choisissent les autres, qui constituent l'équilibre. Nous avons vu que cette connaissance commune permet aussi, en supposant de plus que les agents ont des croyances antérieures communes, d'éviter un problème d'indétermination potentiel dans la notion même de conjecture. Remarquablement absent de ce résultat est la notion de connaissance commune de la rationalité, comme la notion de connaissance commune des préférences. Nous avons observé cependant que la première, la connaissance commune de la rationalité, découle des prémisses de la caractérisation épistémique si l'on y ajoute la seconde, à savoir que les préférences sont aussi connaissance commune. Enfin, rappelons que dans les deux résultats présentés, on y suppose toujours, en arrière-plan, que l'espace d'incertitude lui-même est connaissance commune.

VIII. Discussion générale

Revenons aux deux questions posées en ouverture de ce chapitre au sujet de la connaissance commune. Notre première question portait sur la définition de la connaissance commune, et sur la question de sa réalité psychologique. Notre seconde question portait sur le problème de la nécessité ou pas de la connaissance commune des conventions. Ces deux

questions nous semblent essentielles pour évaluer le lien mis en avant par Lewis entre connaissance commune et convention.

S'agissant de la première question, nous avons vu qu'il existe deux définitions de la connaissance commune, la définition itérative infinitaire d'une part, qui dit qu'une proposition est connaissance commune au sein d'un groupe si chacun la sait, chacun sait que chacun la sait, et ainsi de suite *ad infinitum*, et la définition de Lewis de l'autre, qui dit qu'une proposition est connaissance commune s'il existe une base de connaissance commune qui indique à chacun cette proposition. La notion de base de connaissance commune est par elle-même beaucoup plus restrictive que la définition itérative, en particulier elle n'implique pas que la proposition indiquée par cette base soit connaissance partagée à tous les ordres. Mais elle permet de dériver la définition itérative si l'on suppose, comme le suggère Lewis, que certains des principes de rationalité qui sont au cœur de cette définition sont eux-mêmes susceptibles d'être répliqués indéfiniment par les agents en termes de croyance.

Selon nous, il serait donc incorrect d'opposer la définition de Lewis et la définition itérative. Lewis accepte la définition itérative, mais étant donné une base de connaissance commune, il considère qu'on n'atteint de la connaissance commune pleine au sein d'un groupe qu'à condition que les agents répliquent sans faille les mêmes principes de rationalité. Ce point a son importance, car il permet à Lewis de regarder la définition infinitaire comme une idéalisation. En pratique, Lewis admet que les attentes d'ordre supérieur peuvent être limitées. Par ailleurs, dans *Convention* Lewis admet qu'une base de connaissance commune peut être plus ou moins forte qu'une autre relativement aux attentes qu'elle engendre. Par exemple, Lewis considère que la saillance est une base de connaissance commune « plus faible » que l'accord explicite (1969, p. 57), car selon lui³⁶

« [la saillance] engendre des attentes d'ordre supérieur plus faibles, puisque la saillance d'un équilibre n'est pas une indication très forte que les agents auront tendance à le choisir. »

Cette notion de force relative n'est cependant pas détaillée par Lewis. Deux interprétations sont envisageables. Selon une interprétation, le degré de force d'une base pourrait se mesurer à l'ordre de croyances partagées engendrées par la base au sein du groupe. Selon une autre interprétation, on

³⁶ « [Saliency] (...) generates weaker higher-order expectations, since the saliency of an equilibrium is not a very strong indication that agents will tend to choose it ».

pourrait imaginer que le degré de force affecte non pas la hiérarchie elle-même, mais le degré de croyance associée au premier niveau dans la hiérarchie et aux suivants (par exemple, on pourrait imaginer qu'une base de connaissance commune est plus faible qu'une autre si elle constitue une raison de croire à 70% que P, alors qu'une autre base donnerait une raison de croire à 85% que P)³⁷.

Il est difficile de trancher entre ces deux interprétations, faute d'une formalisation plus explicite dans le texte de Lewis. Néanmoins, ce que montrent ces exemples est que la notion de base de connaissance commune se présente comme une notion plus réaliste que ce que donne la définition itérative courante. Compte tenu des idéalizations faites en théorie des jeux, cependant, la différence entre les deux définitions discutées de la connaissance commune n'est pas significative relativement à la seconde question qui nous intéresse. Pour des agents idéalement rationnels, la définition lewisienne implique en effet la définition itérative infinie, et la question qui nous intéresse est celle de savoir si des croyances d'ordre supérieur de niveau arbitraire sont bel et bien requises pour qu'il y ait coordination entre les agents, et pour qu'il y ait convention en particulier. Quelle leçon pouvons-nous tirer de ce point de vue des résultats d'Aumann et Brandenburger ?

Le premier résultat que nous avons discuté, concernant les équilibres stratégiques dans les jeux à deux joueurs, semble apporter une confirmation de l'idée que la connaissance commune n'est pas une condition nécessaire de la sélection d'un équilibre. En effet, ce résultat dit qu'il suffit que les agents connaissent l'action de l'autre, et maximisent l'utilité espérée, pour qu'il y ait équilibre. Comme le souligne avec éloquence Brandenburger (1992)³⁸.

« Bien que la proposition puisse sembler transparente, elle réserve une surprise : la connaissance commune n'est pas requise. Et ce malgré le folklore qui s'est construit sur la connexion des notions d'équilibre de Nash et de connaissance commune. La base du folklore est à peu près la suivante : le joueur 1 joue son rôle dans l'équilibre de Nash parce qu'il

³⁷ Sur la seconde interprétation, cf. en particulier Monderer et Samet (1989).

³⁸ « *While the proposition may seem transparent, it does contain a surprise : common knowledge is not needed. This is despite the folklore that has grown up connecting the notions of Nash equilibrium and common knowledge. The basis of the folklore goes something like this : player 1 plays his part of a Nash equilibrium because he thinks that player 2 is playing her part, a belief which is held because he thinks she thinks he is playing his part, and so on. The circularity of Nash equilibrium does indeed seem related to the infinite hierarchy of beliefs inherent in common knowledge, yet [the result] suggests that the relationship is illusory.* »

pense que la joueuse 2 joue son rôle, une croyance qu'il a parce qu'il pense qu'elle pense qu'il joue son rôle, et ainsi de suite. La circularité de l'équilibre de Nash semble en effet rattachée à la hiérarchie infinie de croyances inhérente à la connaissance commune, pourtant [le résultat] suggère que la relation entre les deux est illusoire. »

L'interprétation donnée par Brandenburger est cependant fort discutable. Comme le remarquent Osborne et Rubinstein (1994, p. 77), « l'hypothèse que chaque joueur connaît les actions de tous les autres est très forte ». D'ordinaire, nous ne connaissons pas l'action de l'autre au moment où nous devons choisir d'agir. De ce fait, les remarques de Brandenburger sont beaucoup moins déflationnistes qu'elles ne semblent à première vue, puisque l'hypothèse de connaissance du choix de l'autre abolit tout simplement l'incertitude stratégique entre les agents. Néanmoins, Osborne et Rubinstein présentent une généralisation du résultat (au cas de stratégies mixtes) par quoi il suffit pour deux agents que chacun connaisse les conjectures de l'autre et sache que l'autre est rationnel pour que leurs actions soient à l'équilibre (cf. Osborne et Rubinstein (1994), Proposition 78.1). En affaiblissant les conditions, ce résultat étaye de façon plus adéquate la remarque de Brandenburger. Toutefois, on peut encore s'interroger sur ce qui permet à deux agents de connaître les conjectures l'un de l'autre, ou encore de savoir que l'autre est rationnel. Pour Lewis, ne l'oublions pas, les agents dans leurs interactions sont avant tout des « monades sans fenêtre » (*windowless monads*)³⁹. Cela signifie qu'en règle générale, nous n'avons pas accès aux croyances les uns des autres. Le seul accès, dans la perspective lewisienne, serait précisément un événement base de connaissance commune, par quoi chaque agent peut s'attendre à ce que l'autre raisonne de la même manière que soi.

Qu'en est-il du second résultat de Aumann et Brandenburger, pour le cas général ? Cette fois, le résultat fait bel et bien intervenir la connaissance commune, puisqu'il demande que les agents aient connaissance commune de leurs conjectures, et qui plus est, il fait intervenir l'hypothèse de croyances antérieures communes entre les agents. Mais comme nous l'avons souligné, ce résultat ne fait pas pour autant de la connaissance commune (des conjectures) une condition nécessaire de l'équilibre stratégique à proprement parler. Aumann et Brandenburger font en effet valoir⁴⁰:

³⁹ Cf. Lewis 1969, p. 32, où Lewis reprend à dessein l'expression de Leibniz.

⁴⁰ « *It is always possible for the players to blunder into a Nash equilibrium "by accident," so to speak, without anybody knowing much of anything.* »

« Il est toujours possible que les joueurs tombent dans un équilibre de Nash “par accident”, pour ainsi dire, sans que personne ne sache grand chose de quoi que ce soit. »

Cette remarque confirme tout à fait la conception même de Lewis de la sélection des équilibres : un équilibre peut se produire par chance, et même en supposant une population d’agents parfaitement rationnels, ces agents peuvent atteindre un équilibre sans connaître les préférences des uns des autres, ni leurs actions, ni leurs anticipations. Mais Aumann et Brandenburger ajoutent⁴¹:

« Néanmoins, les énoncés sont “stricts”, au sens où il ne peuvent être améliorés ; aucune des conditions ne peut être omise, ou même être affaiblie de façon significative. »

Si les conditions sont strictes à défaut d’être nécessaires, tout le problème est de déterminer quelles autres conditions seraient susceptibles de les rendre pertinentes. Cette remarque invite à la spéculation suivante : ne pourrait-on voir les conditions d’Aumann et Brandenburger comme des conditions susceptibles de s’appliquer, non pas à la sélection d’un premier équilibre, mais à la maintenance d’un équilibre ? Si c’était le cas, on pourrait voir alors dans ce résultat une justification de la position même de Lewis, en vertu de laquelle si la connaissance commune ne joue pas de rôle dans la sélection des équilibres, elle joue un rôle dans leur perpétuation au statut de convention.

Mais admettre qu’elle joue un rôle ne signifie pas qu’elle soit nécessaire. Souvenons-nous que Lewis propose deux définitions de la notion de convention : une première définition qui ne fait pas intervenir la connaissance commune, et une seconde dans laquelle cette clause s’ajoute. Lewis commente la différence précisément sur l’exemple de la conduite automobile. Il imagine la situation suivante⁴² :

⁴¹ « Nevertheless, the statements are “tight,” in the sense that they cannot be improved upon ; none of the conditions can be left out, or even significantly weakened. »

⁴² « Suppose everyone drives on the right because he expects everyone else to drive on the right and he wants to prevent collisions. But suppose no one gives anyone else credit for intelligence equal to his own. Everyone holds this false belief (call it *f*) : “Except for myself, everyone drives on the right by habit, for no reason, and would go on driving on the right no matter what he expected others to do”. This is a case of convention under the unamended definition, despite the false beliefs ; but I think it ought to be excluded. It cannot be a case of convention under the amended definition (unless we are extremely irrational) ».

« Supposons que chacun conduise à droite parce qu’il s’attend à ce que chaque autre conduise à droite et qu’il veut prévenir les collisions. Mais supposons que personne ne fasse crédit à personne d’une intelligence égale à la sienne. Chacun entretient cette croyance fautive (appelons la f) : “excepté pour moi, chacun conduit à droite par habitude, sans raison, et continuerait à conduire à droite quoi qu’il attende de ce que feront les autres”. C’est un cas de convention sous la définition non-amendée, malgré les fausses croyances ; mais je pense qu’il faut l’exclure. Ce ne saurait être un cas de convention sous la définition amendée (à moins que nous ne soyons extrêmement irrationnels). »

Cette situation, notons-le, est très proche du cas qu’imagine Skyrms pour critiquer l’idée que la connaissance commune soit même nécessaire à la perpétuation d’un équilibre. Dans le scénario décrit par Lewis, chaque agent est rationnel mais considère en quelque sorte les autres comme des « bactéries » agissant de façon aveugle. Notons que Lewis n’a pas d’argument direct pour exclure ce genre de situation. D’une certaine manière, Lewis est dans la même position pour la maintenance des équilibres qu’Aumann et Brandenburger pour leur sélection : il ne peut exclure la possibilité que la maintenance des équilibres ne repose, dans certains cas, que sur des attentes de premier ordre des agents, et même sur des croyances fausses concernant les croyances d’autrui. En l’occurrence, chaque agent conduirait à droite parce qu’il croit que les autres le feront inconditionnellement, et qu’il préfère le faire en supposant que les autres eux-mêmes auront fait ce choix.

Qu’est-ce donc qui rend malgré tout pertinente la connaissance commune ? La réponse de Lewis est fondamentalement basée sur la prise en compte comme un fait de la capacité des agents rationnels à se représenter les croyances d’autrui. Pour Lewis, nos conventions sont de facto fondées sur des événements qui constituent des bases de connaissance commune, c’est-à-dire qui engendrent des attentes mutuelles. Plus encore, comme nous l’avons indiqué dans la section 3 plus haut, pour tout jeu de coordination dans lequel les agents n’ont pas des croyances sur les croyances les uns des autres qui soient concordantes à tous les niveaux, on peut construire un scénario par lequel les agents manquent de se coordonner de façon adéquate. Nous pouvons donc interpréter le rôle de la connaissance commune pour les conventions de la même façon que Aumann et Brandenburger le font pour les équilibres plus généralement : comme une condition certes non-nécessaire, mais comme une condition suffisante stricte. De fait, pour Lewis, dès lors que des agents rationnels sont des

agents qui se représentent les croyances les uns des autres, la connaissance commune est ce qui garantit la fiabilité de la coordination. L'illustration la plus claire de cette interprétation en est donnée par le commentaire que fait Lewis (1973) de la pertinence de la clause de connaissance commune pour sa définition d'une convention⁴³ :

« Cette condition assure la stabilité. Si quelqu'un essaie de répliquer le raisonnement d'un autre, peut-être en incluant la réplification de l'autre de son raisonnement, ... , le résultat renforcera plutôt qu'il ne subvertira son attente de conformité à [la régularité] R. Peut-être une version négative de (6) [la condition de connaissance commune] ferait-elle l'affaire : personne ne manque de croire que (1) à (5) [les autres conditions définissant la convention] valent, personne ne croit que les autres manquent de croire ceci, et ainsi de suite. »

Comme l'indique cette citation, le rôle de la connaissance commune est donc mal compris dans la perspective lewisienne si l'on y cherche une condition de réalisation de la coordination. Dans certains cas, certes, la connaissance commune peut être condition suffisante de la convergence à l'équilibre. Mais dans le cas général, Lewis conçoit la notion comme graduée, et comme qualifiant la force des attentes : le degré de connaissance commune effectivement atteint n'assure pas de la réalisation d'un équilibre, mais de la propension plus ou moins forte à le répliquer de telle ou telle manière.

IX. Conclusion

A quoi donc sert la connaissance commune pour la coordination entre les agents ? La leçon que nous tirons de ce chapitre est que la connaissance commune n'est à proprement parler ni une condition nécessaire de sélection des équilibres, ni une condition nécessaire de leur maintenance. A première vue, le déflationnisme de Skyrms touchant la pertinence de la connaissance commune semble devoir triompher. Nous faut-il alors employer le rasoir d'Occam et réputer inutile le concept inventé par Lewis ?

La réponse à cette question nous paraît négative. Tout d'abord, comme l'indique distinctement le texte de *Convention*, Lewis était conscient

⁴³ « *This condition ensures stability. If anyone tries to replicate another's reasoning, perhaps including the other's replication of his reasoning, ..., the result will reinforce rather than subvert his expectation of conformity to R. Perhaps a negative version of (6) would do the job : no one disbelieves that (1) to (5) hold, no one believes that the others disbelieve this, and so on.* ».

de ces deux objections au caractère nécessaire de la connaissance commune. S'agissant de l'occurrence d'un équilibre, Lewis admet que celui-ci peut se produire par chance. S'agissant de sa maintenance au statut d'une régularité de comportement, il admet également que l'équilibre peut théoriquement se maintenir sur la base de croyances fausses dans le comportement routinier des agents. Mais le fait que la connaissance commune ne soit pas une condition nécessaire de l'émergence ni même de la perpétuation d'un équilibre n'invalide pas pour autant sa thèse centrale, à savoir que la connaissance commune joue un rôle déterminant dans la manière dont un premier équilibre de coordination acquiert le statut d'une convention chez des agents rationnels. Pour le dire autrement, la connaissance commune, fondée chez Lewis sur la notion d'événement-base de connaissance commune, est un élément sans lequel une régularité de comportement manquera d'être répliquée de façon fiable chez des agents doués d'une théorie de l'esprit, c'est-à-dire capables de se représenter les raisons d'agir les uns des autres. La caractérisation formelle de l'équilibre de Nash que donnent Aumann et Brandenburger, qui laisse de côté les considérations empiriques sur lesquelles se fonde Lewis, en décrivant des agents parfaitement rationnels, nous apparaît comme une confirmation indirecte de ce point de vue. Cette caractérisation, comme nous l'avons vu, s'appuie intégralement sur la définition itérative de la connaissance commune, qu'accepte Lewis. Néanmoins, la définition lewisienne de la connaissance commune est dans une large mesure plus ambitieuse que la définition itérative, puisque la notion de base de connaissance commune vise à rendre compte à la fois de l'intensité et de l'émergence des attentes réciproques chez les agents rationnels. Cette notion d'émergence, thématifiée chez Lewis dans les notions d'accord, de préférence et de saillance empruntées à Schelling, reste encore aujourd'hui un paramètre exogène à la plupart des modèles formels de théorie des jeux.

Remerciements

Nous remercions Yann Schmitt de son invitation, ainsi que le Département d'Études Cognitives de l'ENS et la Société de Philosophie Analytique qui ont rendu possible notre collaboration. Nos remerciements vont également à Cédric Paternotte et à un rapporteur anonyme de la revue *Klesis* pour leurs commentaires sur la première version de ce travail.

Paul Egré : Je profite de cet essai pour rendre hommage à David Lewis, qui fut l'un de mes professeurs à Princeton en 1998-1999. C'est grâce à un échange entre l'ENS et Princeton que je pus passer une année comme *Visiting graduate student* en qualité de Jane Eliza Procter Fellow au département de philosophie de l'université de Princeton. Aucun des cours que donnait Lewis cette année-là n'était consacré à la connaissance commune ou à la théorie de la rationalité (Lewis à cette période s'intéressait beaucoup plus à la métaphysique, il enseignait cette année-là un cours sur le voyage dans le temps, un cours sur la causalité, et un cours de philosophie des sciences). Mais *Convention* est le premier texte que j'y ai lu (encouragé par Jacques Dubucs, auquel va également ma gratitude), et la connaissance commune fut notamment l'objet d'un bref entretien avec David Lewis lors de ce séjour. Mon souvenir de cette discussion est aujourd'hui confus, mais suffisant pour me rappeler qu'il portait sur la question de savoir si le concept de connaissance commune ne demandait pas un traitement probabiliste. Dans mon souvenir, Lewis jugeait inutile ce raffinement, estimant sa définition originale suffisamment réaliste.

Olivier Roy : Les idées développées ici ont profité de nombreuses discussions avec Eric Pacuit. Je suis également reconnaissant à Robert Aumann de m'avoir accordé l'entretien paru dans Hendricks et Roy (2010), et à Vincent F. Hendricks pour son soutien dans la réalisation de cet entretien. Ce travail fut rendu possible par le soutien financier de la fondation Alexander von Humboldt.

Bibliographie

- R. Aumann, « Agreeing to disagree ». *Annals of Statistics*, 4, p. 1236–1239, 1976.
- R. Aumann, « Correlated equilibrium as an expression of Bayesian rationality », *Econometrica*, 55(1) : 1–18, 1987.
- R. Aumann et A. Brandenburger. « Epistemic conditions for Nash equilibrium », *Econometrica*, 63(5) : 1161–1180, 1995.
- J. Barwise, « Three views of common knowledge », in *Proceedings of the 2nd Conference on Theoretical Aspects of Reasoning about Knowledge*, pp. 365–379, Morgan Kaufmann Publishers Inc, 1988.
- G. Bonanno et K. Nehring, « Agreeing to disagree, a survey », 1997, Manuscrit disponible sur le site web des auteurs.

- G. Bonanno et K. Nehring, « How to make sense of the common prior assumption under incomplete information », *International Journal of Game Theory*, Vol.26, No.3, 1999.
- A. Brandenburger, « Knowledge and Equilibrium in Games », *The Journal of Economic Perspectives*, Vol. 6, No. 4 pp. 83-101, 1992.
- H. H. Clark et C. R. Marshall, « Definite reference and mutual knowledge ». In A. K. Joshi, B. Webber, & I. Sag (Eds.), *Elements of discourse understanding* (p. 10-63). Cambridge : Cambridge University Press, 1981.
- V. Crawford et J. Sobel, « Strategic information transmission », *Econometrica* 50, p. 1431–1451, 1982.
- R. P. Cubitt et R. Sugden, « Common knowledge, salience and convention : a reconstruction of David Lewis' game theory », *Economics and Philosophy*, 19 : 175–210, 2003.
- R. Fagin, J. Halpern, Y. Moses, M. Vardi, *Reasoning about knowledge*. MIT Press, 1995.
- H. P. Grice, « Meaning ». *The Philosophical Review* 66 (3) : 377-388, 1957
- J. Halpern, « Characterizing the Common Prior Assumption », *Journal of Economic Theory*, vol.106, no.2, 2002.
- J.C. Harsanyi, « Games with incomplete information played by 'bayesian' players, I - III », *Management Science*, 14 : p. 159–182, 320–334, 486–502, 1967-68.
- A. Heifetz, « The positive foundation of the common prior assumption », *Games and Economic Behavior*, vol. 56, no.1, 2006.
- V. F. Hendricks et O. Roy, *Epistemic logic : 5 questions*. Automatic Press /VIP, 2010.
- D. Lewis, *Convention. A philosophical study*, Basil Blackwell, 1969.
- D. Lewis, « Psychophysical and theoretical identifications » (1972) *The Australasian Journal of Philosophy*, 50, p. 249–258. Repr. in D. Lewis, *Papers in Metaphysics and Epistemology*, 1999, chap. 16, p. 248–261
- D. Lewis, « Language and Languages » (1973), *Versus* 4, pp. 2-21, repr. in D. Lewis, *Philosophical Papers*, 1983, vol. 1, chap. 11, pp. 163-188.
- D. Monderer & D. Samet, « Approximating common knowledge with common beliefs », *Games and Economic Behavior* 1 (2), p. 170–190, 1981.

- S. Morris, « The common prior assumption in economic theory », *Economics and Philosophy*, vol. 11, 1995.
- R. Meyerson, *Game theory : analysis of conflict*, Harvard University Press, 1997.
- C. Osborne et A. Rubinstein, *A course in game theory*. MIT Press, 1994.
- C. Paternotte, « Being realistic about common knowledge : a Lewisian approach », *Synthese* 183, p. 249–276, 2011.
- A. Perea, *Epistemic Game Theory : Reasoning and Choice*. Cambridge University Press, 2012.
- B. Polack, « Epistemic conditions for Nash equilibrium, and common knowledge of rationality », *Econometrica*, vol.67, no.3, 1999
- D. Rosthchild, « From convention to implicature ». Manuscrit, Oxford University, 2012.
- T. Schelling, *The strategy of conflict*, Harvard University Press, 1960.
- S. Schiffer, *Meaning*. Oxford University Press, 1972.
- B. Skyrms, *Evolution of the social contract*, Cambridge University Press, 1996.
- B. Skyrms, *The stag hunt and the evolution of social structure*, Cambridge University Press, 2004.
- B. Skyrms, *Signals. Evolution, learning, & information*, Oxford University Press, 2010.
- G. Sillari, « A logical framework for convention ». *Synthese*, 147(2), p. 379–400, 2005.
- M. Spence, « Job market signaling », *The Quarterly Journal of Economics*, 87 (3), p. 355–374, 1973.
- P. Vanderschraaf et G. Sillari, « Common knowledge ». In Zalta, E. N., (éd.), *The Stanford Encyclopedia of Philosophy*. Spring 2009 edition.